Perform regression analysis on breast cancer

Group members: Kexin Fan, Xinyi Xie, Jiawei Zeng

Introduction

Breast cancer is one of the most common cancers diagnosed in U.S. women. Many of us think of breast cancer as a female disease, but it can also occur in men. Breast cancer is found when the cells in the breast begin to grow uncontrollably. We filtered out some of the important features in the progress of searching data, compared and performed regression on breast cancer. Based on these features, we can use regression modeling to determine whether a tumor is benign or malignant.

Data:

Attribute Information:
1)ID number
2) Diagnosis (M = malignant, B = benign)
3-32)

Ten real-valued features are computed for each cell nucleus:
a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

According to these data, we can use these methods to predict the tumor is whether benign or malignant.

Methods

- Exploratory Data Analysis
- Multiple regression model
- Principal component analysis
- Lasso Regression
- K-fold cross validation
- Linear Discriminant Analysis
- Decision tree

Goals

The aim of this project was to develop a regression model that could predict whether a tumor is benign or malignant based on the given characteristics in our dataset. And study how to improve the performance of the algorithm for the original dataset. And how much the various characteristics of a tumor influence whether it is ultimately benign or malignant.

Dataset

https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?select=data.csv
https://www.kaggle.com/datasets/vijayaadithyanvg/breast-cancer-prediction?select=data.csv